**Remarks**

Claims 1-13 are pending in the application. Claims 1-13 are rejected. All rejections are respectfully traversed.

The invention provides a method for determining similarities of interpretation between portions of multimedia (videos) at a very high level, e.g., similar action in an adventure movie, scoring opportunities in a sports video, romantic activity in a gothic movie, fright in a horror movie, humor in a comedy movie, and so forth. The term 'high-level' is used because the similarity considers a sequence of semantic events extended over a relatively long time period.

Therefore, the invention segments multimedia content to extract video object planes, which can encode arbitrary-shaped objects according to the MPEG-4 standard, also known as H.264 or AVC, see Specification, page 2:

> "Newer video coding standards, such as MPEG-4, see "Information Technology -- Generic coding of audio/visual objects," ISO/IEC FDIS 14496-2 (MPEG-4 Visual), Nov. 1998, allow arbitrary-shaped objects to be encoded and decoded as separate video object planes (VOP)... The most recent standardization effort taken on by the MPEG committee is that of MPEG-7, formally called "Multimedia Content Description Interface," see "MPEG-7 Context, Objectives and Technical Roadmap," ISO/IEC N2729, March 1999. Essentially, this standard plans to incorporate a set of descriptors and description

schemes that can be used to describe various types of multimedia content."

In the art, these newer, high-level structures are distinguished from older, low-level features such as color and motion.

Claims 1-13 are rejected under 35 U.S.C. 103(a) as being unpatentable over Yeo et al., U.S. Patent No. 5,821,945 (Yeo), in view of Boetje et al., U.S. Patent No. 6,049,332 (Boetje).

Video object planes (VOP) are defined in the H.264/MPEG-4 or AVC standards. The call for proposals was in May 1998, and the first draft design for the new standard was not adopted in until 1999. The Yeo patent application was filed in May 1997. Yeo could not have known about video object planes as claimed.

The invention segments multimedia content to extract video object planes. The decomposition of videos into a hierarchical scene transition graph according to Yeo reflects acts, scenes and shots of the video, not video object planes.

Yeo does not extract and associate features of the video object planes to produce content entities. Instead, the browsing process of Yeo is "automated to extract a hierarchical decomposition of a complex video selection in four steps: the identification of video shots, the clustering of video shots of similar visual contents, the presentation of the content and structure to the

users via the scene transition graph, and finally the hierarchical organization of the graph structure."

Yeo does not measure high-level attributes of each content entity. Yeo states:

> "Low level vision analyses operated on video frames achieve reasonably good results for the measurement of similarity (or dissimilarity) of different shots. Similarity measures based on image attributes such as color, spatial correlation and shape can distinguish different shots to a significant degree, even when operated on much reduced images as the DC images. Both color and simple shape information are used to measure similarity of the shots."

The problems with low-level features as in Yeo are distinguished in the present application at pages 2 and 3:

> "Another problem with such low-level descriptors, in general, is that a high-level interpretation of the object or multimedia content is difficult to obtain. Hence, there is a limitation in the level of representation. To overcome the drawbacks mentioned above and obtain a higher-level of representation, one may consider more elaborate description schemes that combine several low-level descriptors. In fact, these description schemes may even contain other description schemes, see "MPEG-7 Description Schemes (v0.5)," ISO/IEC N2844, July 1999."

Yeo does not describe content entities and comparing the ordered content entities in a plurality of the directed acyclic graphs to determine similar interpretations of the multimedia content.

Boetje does not teach comparing ordered content entities in a plurality of directed acyclic graphs (DAG). The Boetje system and method "creates a broadcast tree comprising a hierarchy of broadcast constituents, each constituent represented as a node in the tree," see abstract and summary. The hierarchy is necessary to traverse the tree in an up and down order. "Thus, to generate a broadcast, the tree is traversed beginning at the highest order constituent, and for each higher order constituent, the associations among lower order constituents of the same order are evaluated to determine the sequence the lower order constituents are to be played." Figures 9 and 10 show hierarchical trees not DAG, as claimed and known in the art.

A DAG is a directed graph with no directed cycles. Every directed acyclic graph corresponds to a partial order on its vertices. A hierarchical tree as in Boetje is a complete ordering of the vertices. The Boetje tree is to automatically detect inconsistencies in a schedule programmed by a programmer, and not to determine similar interpretations of multimedia content as claimed, see:
***(col. 33, line 40-col. 34, line 62; and see figs. 22a-b and the associated text).***

The invention measures attributes of content entities that include intensity attributes. Yeo, at column 7, lines 35, et seq., measures *correlations* between *frames* as differences:

The inventors discovered that measuring correlation between two small images (even the DC images) does give a very good indication of similarity (it is actually dissimilarity in the definition below) in these images. By using the sum of absolute difference, the correlation between two images, $f_m$ and $f_n$ is commonly computed by:

$$\epsilon(m,n) = \sum_{j=1}^{J} \sum_{k=1}^{K} |f_m(j,k) - f_n(j,k)|. \quad (8)$$

Yeo does not measure attributes of content entities that include direction attributes. Yeo measures, in images, the "two-dimensional moment invariant of the luminance." Those of ordinary skill in the art would not confuse direction and luminance, see column 7, lines 13-19.

Yeo does not measure attributes of content entities that include spatial attributes and the order is spatial. The Yeo measurements take place on frames.

Yeo does a temporal segmentation, but never measures temporal attributes of content entities.

Yeo does not rank order attributes measured of content entities.

The scene transition graph in Yeo is not derived form video object planes.

There is nothing in column 19 that would indicate that Yeo generates a summary of a video.

The user is given the flexibility to interactively select the number of clusters desired or to set caps on the dissimilarity values between individual shots allowed in a cluster. In test trials of the present system, after the initial shot partitions, the user only needs to slightly adjust the knobs to change these partitions to yield satisfactory results, often with less than four such trials. FIGS. 3*a* and 3*b* show clustering results on two sequences: a 16-minute Democratic Convention video sequence, and a News Report, respectively. The directed scene transition graph is laid out using the algorithms disclosed by László Szirmay-Kalos, in a paper "Dynamic layout algorithm to display general graphs", in *Graphics Gems IV*, pp. 505–517, Academic Press, Boston, 1994. FIGS. **4** and **5** show the sample interface and graph layout of the two above-mentioned video sequences, based on the results in FIGS. 3*a* and 3*b*, respectively. Each node represents a collection of shots, clustered by the method described above. For simplicity, only one frame is used to represent the collection of shots. A means is also provided for the users to re-arrange the nodes, group the nodes together to form further clusters, and ungroup some shots from a cluster. This enables the user to organize the graphs differently to get a better understanding of the overall structures.

Columns 6 though 8 also do not describe a video summary. Applicants respectfully request the Examiner to point out which word(s) in Yeo mean "a video summary." The Applicants have carefully read Yeo but cannot find any video summarization steps.

At column 7, Yeo states:

Shape

The present system uses as another measure of similarity between two images the two-dimensional moment invariant of the luminance. However, the inventors discovered that the order of magnitudes in different moment invariant vary greatly: in many examples the ratio of first moment invariant to the third or fourth moment invariant can vary by several orders of magnitude.

By using the Euclidean distance of the respective moment

There is absolutely nothing there that would suggest that a measure of similarity between two-dimensional luminance would suggest a three dimensional video. A three dimensional video is a video that also includes depth, such as a MRI video or CAT scan.

Claimed are directed acyclic graphs where nodes represent the content entities and edges represent breaks in the segmentation, and the measured attributes are associated with the corresponding edges. Yeo teaches a graph "with nodes representing scenes and edges representing the progress of the story from one scene to the next."

Claimed is at least one secondary content entity associated with a particular content entity, and wherein the secondary content entity is selected during the traversing. Nowhere in columns 2-6 are these limitations described.

Claimed is a summary of the multimedia with a selected permutation of the content entities according to the associated ranks. At columns 9:

> The user is given the flexibility to interactively select the number of clusters desired or to set caps on the dissimilarity values between individual shots allowed in a cluster. In test trials of the present system, after the initial shot partitions, the user only needs to slightly adjust the knobs to change these partitions to yield satisfactory results, often with less than four such trials. FIGS. **3***a* and **3***b* show clustering results on two sequences: a 16-minute Democratic Convention video sequence, and a News Report, respectively. The directed scene transition graph is laid out using the algorithms disclosed by László Szirmay-Kalos, in a paper "Dynamic layout algorithm to display general graphs", in *Graphics Gems IV*, pp. 505–517, Academic Press, Boston, 1994. FIGS. **4** and **5** show the sample interface and graph layout of the two above-mentioned video sequences, based on the results in FIGS. **3***a* and **3***b*, respectively. Each node represents a collection of shots, clustered by the method described above. For simplicity, only one frame is used to represent the collection of shots. A means is also provided for the users to re-arrange the nodes, group the nodes together to form further clusters, and ungroup some shots from a cluster. This enables the user to organize the graphs differently to get a better understanding of the overall structures.

Yeo allows the user to rearrange nodes in a graph. There is nothing there to indicate that content entities can be permuted according to ***rank***.

It is believed that this application is now in condition for allowance. A notice to this effect is respectfully requested. Should further questions arise concerning this application, the Examiner is invited to call Applicants' attorney at the number listed below. Please charge any shortage in fees due in connection with the filing of this paper to Deposit Account 50-0749.

Respectfully submitted,
Mitsubishi Electric Research Laboratories, Inc.

By
/Dirk Brinkman/

Dirk Brinkman
Attorney for the Assignee
Reg. No. 35,460

201 Broadway, 8th Floor
Cambridge, MA 02139
Telephone: (617) 621-7539
Customer No. 022199